On the Finite-Time Complexity and Practical Computation of Approximate Stationarity Concepts of Lipschitz Functions

Lai Tian 1  Kaiwen Zhou 2  Anthony Man-Cho So 1

Abstract

We report a practical finite-time algorithmic scheme to compute approximately stationary points for nonconvex nonsmooth Lipschitz functions. In particular, we are interested in two kinds of approximate stationarity notions for nonconvex nonsmooth problems, i.e., Goldstein approximate stationarity (GAS) and near-approximate stationarity (NAS). For GAS, our scheme removes the unrealistic subgradient selection oracle assumption in (Zhang et al., 2020, Assumption 2) and computes GAS with the same finite-time complexity. For NAS, Davis & Drusvyatskiy (2019) showed that ρ-weakly convex functions admit finite-time computation, while Tian & So (2021) provided the matching impossibility results of dimension-free finite-time complexity for first-order methods. Complement to these developments, in this paper, we isolate a new class of functions that could be Clarke irregular (and thus not weakly convex anymore) and show that our new algorithmic scheme can compute NAS points for functions in that class within finite time. To demonstrate the wide applicability of our new theoretical framework, we show that ρ-margin SVM, 1-layer, and 2-layer ReLU neural networks, all being Clarke irregular, satisfy our new conditions.

1. Introduction

In this paper, we consider the following general optimization problem for an $L$-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$

$$\min_{x \in \mathbb{R}^d} f(x), \quad (\triangledown)$$

where $f$ could be both nonsmooth and nonconvex (“non”-setting for short). We are particularly interested in algorithms with a finite-time complexity for computing approximately stationary points of Problem (\triangledown). Note that when $f$ is smooth, it is folkloric that computing an $\epsilon$-stationary point (i.e., $\|\nabla f(x)\| \leq \epsilon$) only requires $O(\epsilon^{-2})$ calls, which is dimension-independent and finite, to the gradient oracle with gradient descent (Nemirovskij & Yudin, 1983).

In the general Lipschitz “non”-setting, a widely used generalized subdifferential $\partial f(x)$ is due to Clarke (1990, Section 2.1) (see also Definition 2.1), which reduces to the convex subdifferential (resp. gradient) if $f$ is convex (resp. smooth). Therefore, by mimicking results in the smooth scenario, it is natural to conjecture that we may be able to design algorithms to compute elements in $\{x : \text{dist}(0, \partial f(x)) \leq \epsilon\}$ in finite time with high probability. However, as shown by Zhang et al. (2020, Theorem 5), that is impossible for any first-order method. Thus, it is curious to ask: What kind of approximate stationarity concept in the “non”-setting will admit dimension-free finite-time computation?

Davis & Drusvyatskiy (2019) gave a nice answer for the class of ρ-weakly convex functions1 by introducing a notion named near-approximate stationarity (NAS, see Definition 2.5), which is closely related to the gradient of the Moreau envelope of $f$. They showed that a subgradient-type method computes an $(\epsilon, \delta)$-NAS point with $O(\rho^2 \delta^{-4} + \epsilon^{-4})$ calls to the subgradient oracle. However, many modern ML models are indeed not weakly convex, e.g., neural networks with ReLU activation functions. Even worse, by extending the Lipschitz hardness results in (Kornowski & Shamir, 2021), Tian & So (2021) demonstrated that, for any finite $T$, there exists a finite $\rho(T)$ such that, for any $0 \leq \epsilon, \delta < \frac{1}{2}$ uniformly, computing an $(\epsilon, \delta)$-NAS point for $\rho(T)$-weakly convex functions within $T$ steps is impossible.

On the other front, starting from the seminal work of Goldstein (1977), a notion named Goldstein approximate stationarity (GAS, see Definition 2.1) exhibits favorable algorithmic consequences. The story begins with an approximation

1 Recall $f$ is ρ-weakly convex if $f(x) + \frac{\rho}{2} \|x\|^2$ is convex. Weak convexity implies Clarke regularity (Vial, 1983, Proposition 4.5).
2 $g(x) = -\max\{x, 0\}$ is not Clarke regular (cf. (Clarke, 1990, Definition 2.3.4)) and not ρ-weakly convex for any $\rho \in \mathbb{R}$.
of the Clarke subdifferential $\partial_f(x)$ (see Definition 2.2). If we update iteratively with
\[\begin{align*}
x_{k+1} & \leftarrow x_k - \delta \cdot g_k/\|g_k\|,
\end{align*}\]
where $g_k := \arg\min_{g \in \partial f(x_k)} \|g\|$ is the minimal norm element in $\partial f(x_k)$, then we can compute an $(\epsilon, \delta)$-GAS point in $O(\epsilon^{-1}\delta^{-1})$ steps. The problem is that obtaining $g_k$ for a general Lipschitz function can be computationally expensive (if possible at all) as there is no known approach to compute $\partial f(x)$. Thus, it is of interest to ask (Kornowski & Shamir, 2021, Proposition 1) and convex (if possible at all) as there is no known approach to compute Definitions 2.4 and 2.5.

1.1. Prior Arts

Asymptotic Analysis. The asymptotic computability of Clarke stationary points (i.e., $\{x : 0 \in \partial f(x)\}$) has been well-understood for quite general functions. With a differential inclusion perspective, Benaim et al. (2005); Majewski et al. (2018); Davis et al. (2020) studied the asymptotic convergence of subgradient-type methods. In particular, Davis et al. (2020) proved the asymptotic convergence to Clarke stationary points for Whitney stratifiable objective functions, which include deep ReLU neural networks as a special case. Daniilidis & Drusvyatskiy (2020) demonstrated that the vanilla subgradient method may not converge for general Lipschitz functions even in continuous time.

Finite-Time Analysis. In contrast to the asymptotic regime, the finite-time complexity in the general “non”-setting is still developing. On the positive side, Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019 showed that for $\rho$-weakly convex functions, $(\epsilon, \delta)$-NAS is computable with $O(\rho^4\delta^{-4} + \epsilon^{-4})$ oracle calls. On the negative side, Kornowski & Shamir (2021) showed that computing NAS for Lipschitz functions in dimension-independent finite time is impossible. Tian & So (2021) sharpened the hardness results for NAS to $\rho$-weakly convex with unbounded $\rho$, thus matching the positive results. For GAS, the gradient sampling scheme (Burke et al., 2005; Kiwiel, 2007; Burke et al., 2020) promises finite but dimension-dependent complexity. (Zhang et al., 2020) reported a novel dimension-independent finite-time algorithm with a impractical subgradient oracle. A recent concurrent work (Davis et al., 2021) adopted similar strategy as our Section 3.2 with different algorithmic implementation. Another line of research is to exploit structure: Duchi & Ruan (2018); Drusvyatskiy & Paquette (2019); Davis & Drusvyatskiy (2019); Bolte et al. (2018); Beck & Hallak (2020). In these settings, nonsmoothness and nonconvexity are properly separated making finite-time analysis possible.

1.2. Contributions

We highlight the main contributions as follows.

- For **Q1**, we report a practical algorithmic scheme to compute GAS points for general Lipschitz functions with finite-time complexity in both deterministic and stochastic settings.
- For **Q2**, we isolate a new function class within which our new algorithmic scheme computes NAS points in finite time. This goes far beyond existing $\rho$-weakly convex results. Besides, we establish a series of theoretical tools to compute parameters in our new function class.
- To demonstrate the wide applicability of the new theoretical framework, we show that $\rho$-margin SVM, 1-layer, and 2-layer ReLU neural networks, all being Clarke irregular, satisfy our new conditions.
2. Preliminaries

In this section, we introduce the necessary background on variational analysis for Lipschitz functions. To begin, we recall the following definition of Clarke subdifferential (Rockafellar & Wets, 2009, Theorem 9.61).

**Definition 2.1** (Clarke subdifferential). Given a point $x$, the Clarke subdifferential of Lipschitz $f$ at $x$ is defined by

\[
\partial f(x) := \{ s : \exists x' \to x, \nabla f(x') \text{ exists}, \nabla f(x') \to s \}.
\]

The following $\delta$-approximation of Clarke subdifferential introduced by Goldstein (1977) has nice theoretical properties and is convenient for algorithmic developments.

**Definition 2.2** (Goldstein $\delta$-subdifferential). Given a point $x$ and $\delta \geq 0$, the Goldstein $\delta$-subdifferential of Lipschitz $f$ at $x$ is defined by

\[
\partial^{\delta} f(x) := \{ s \in \partial f(x) : \| s \| \leq \delta \}.
\]

We record some useful properties of the Clarke subdifferential and its Goldstein approximation here:

**Fact 2.3** (cf. Clarke (1990); Goldstein (1977); Zhang et al. (2020)). For an $L$-Lipschitz continuous $f$ and $\delta > 0$,

- $\partial f(x), \partial^{\delta} f(x)$ are nonempty, convex, compact;
- $\partial f(x) = \bigcap_{\delta > 0} \partial^{\delta} f(x)$;
- $\partial^{\delta} f(x) = \bigcap_{\delta > 0} \partial f(x)$;
- if $f$ is $C^2$ near $x$, then $\partial f(x) = \{ \nabla f(x) \}$;
- if $f$ is convex, then $\partial f(x)$ is the convex subdifferential.

We are now ready to introduce two important approximate stationarity notions. We refer the reader to (Davis & Drusvyatskiy, 2020) for a nice expository material.

**Definition 2.4** (Goldstein approximate stationarity, GAS). Given a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we say that $x \in \mathbb{R}^d$ is an $(\epsilon, \delta)$-GAS point if

\[
\text{dist}
\left(
0, \partial^{\delta} f(x)
\right)
\leq \epsilon.
\]

**Definition 2.5** (near-approximate stationarity, NAS). Given a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we say that $x \in \mathbb{R}^d$ is an $(\epsilon, \delta)$-NAS point if

\[
\text{dist}
\left(
0, \bigcup_{y \in \partial f(x)} \partial f(y)
\right)
\leq \epsilon.
\]

It is easy to see that if $x$ is NAS, then $x$ is also GAS as $\partial f(x) \supseteq \bigcup_{y \in \partial f(x)} \partial f(y)$. But the converse does not hold in general, even for continuously differentiable functions.

**Fact 2.6** (Kornowski & Shamir (2021, Proposition 1)). For any $\delta > 0$, there exists a continuously differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$, which is $2\pi$-Lipschitz on $\partial \mathbb{B}$, such that $(0, 0)$ is $(0, \delta)$-GAS but $\min_{y \in \partial \mathbb{B}} \| \nabla f(y) \| \geq 1$.

Fact 2.6 does not hold for $p$-weakly convex functions with sufficiently small $\delta$. Thus, it is still unclear whether NAS and GAS are equivalent assuming $\rho$-weak convexity with finite $\rho \geq 0$. We report below a convex polyhedral version (recall that convexity is $0$-weak convexity), which might be of independent interest.

**Proposition 2.7** (convex polyhedron). For any $\delta > 0$, there exists a convex function $f : \mathbb{R}^2 \to \mathbb{R}$, which is $2$-Lipschitz with polyhedral $\partial f$, such that $(0, 2\delta)$ is $(0, \delta)$-GAS but $\min_{y \in \partial \mathbb{B}(0, 2\delta)} \text{dist}(0, \partial f(y)) \geq \frac{2}{5} \sqrt{\delta}$.

3. Computing GAS with Practical Oracle by Random Conic Perturbation

3.1. Subgradient Oracles

**Assumption 3.1** (practical oracle). Given $x$ and Lipschitz continuous $f$:

(a) In the deterministic setting, if $f$ is differentiable at $x$, then the oracle $O_d(x)$ returns a function value $f(x)$ and the gradient $\nabla f(x)$. Otherwise, it sets error $= 1$.

(b) In the stochastic setting, if $f$ is differentiable at $x$, then the oracle $O_s(x)$ returns a stochastic gradient $g_x$ with $\mathbb{E}[g_x] = \nabla f(x)$ satisfying $\mathbb{E}[\| g_x - \nabla f(x) \|^2 | \sigma(x)] \leq \sigma^2$. Otherwise, it sets error $= 1$.

Compared with the oracle in (Zhang et al., 2020, Assumption 1), Assumption 3.1 only needs to evaluate the gradient $\nabla f(x)$ at differentiable points. Indeed, many modern Automatic Differentiation software (e.g., PyTorch, TensorFlow) can be used as an implementation of Assumption 3.1 without worrying about their incorrectness on subgradient evaluation for nonconvex nonsmooth function (Kakade & Lee, 2018).

3.2. Deterministic Setting

In this section, we present a practical algorithm for computing an $(\epsilon, \delta)$-GAS point and establish its finite-time complexity in the deterministic setting. The new algorithm replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(a)) with Assumption 3.1(a).

3.2.1. ALGORITHM

The main idea is to make use of the almost everywhere differentiability of Lipschitz functions as guaranteed by
Rademacher’s Theorem. By adopting a random conic perturbation to the uniform sampling direction in (Zhang et al., 2020, Algorithm 1), we have the following Algorithm 1, in which the main differences with (Zhang et al., 2020, Algorithm 1) are marked in blue. See also Figure 1.

Algorithm 1 Perturbed INGD

Input: \( x_1 \in \mathbb{R}^d \).
1. Set \( \epsilon_{\text{error}} = 0 \).
2. for \( t \in [T] \) do
3. \hspace{1em} while \( \|m_{t,k}\| > \epsilon \) do
4. \hspace{2em} Sample \( y_{t,1} \) uniformly from \( \mathbb{B}_\delta(x_t) \).
5. \hspace{2em} Call oracle \( \sim, m_{t,1} = \mathcal{O}_d(y_{t,1}) \).
6. \hspace{2em} for \( k \in [K] \) do
7. \hspace{3em} \( x_{t,k} = x_t - \left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta \frac{m_{t,k}}{\|m_{t,k}\|} \).
8. \hspace{3em} if \( \|m_{t,k}\| \leq \epsilon \) then
9. \hspace{4em} Terminate the algorithm and return \( x_t \).
10. \hspace{3em} else if \( f(x_{t,k}) - f(x_t) < -\frac{\delta}{2} \|m_{t,k}\| \) then
11. \hspace{4em} Set \( x_{t+1} = x_{t,k} \) and \( t = t + 1 \).
12. \hspace{3em} Break while-loop.
13. \hspace{3em} else
14. \hspace{4em} Sample \( u_{t,k+1} \in \mathbb{R}^{d+1} \) uniformly from \( \mathbb{S}^d \).
15. \hspace{4em} Let \( v_{t,k+1} \) be the first \( d \) coordinates of \( u_{t,k+1} \).
16. \hspace{4em} \( b_{t,k+1} = v_{t,k+1} - m_{t,k} \cdot \frac{x_{t,k} - x_{t,k-1}}{\|x_{t,k} - x_{t,k-1}\|} \cdot (x_t - x_{t,k}) \).
17. \hspace{4em} Sample \( y_{t,k+1} \) uniformly from \( \mathbb{B}_\delta(x_t, x'_{t,k}) \), where \( x'_{t,k} := x_{t,k} + \frac{m_{t,k}}{8L} \cdot b_{t,k+1} \).
18. \hspace{4em} Call oracle \( \sim, g_{t,k+1} = \mathcal{O}_d(y_{t,k+1}) \).
19. \hspace{4em} Update \( m_{t,k+1} = \beta_{t,k} m_{t,k} + (1 - \beta_{t,k}) g_{t,k+1} \) with \( \beta_{t,k} := \frac{8L^3 - 4L \|m_{t,k}\|}{8L^3 - 4L \|m_{t,k}\| - 4L \|m_{t,k}\|^2} \).
20. \hspace{3em} end if
21. \hspace{2em} end for
22. \hspace{1em} end while
23. end for

3.2.2. Finite-Time Analysis

The main technical contributions in the analysis are summarized in the following two lemmas.

Lemma 3.2. Let \( D := \{x : f \text{ is differentiable at } x\} \). Given locally Lipschitz continuous \( f \), we have

\[ P \left( \exists (t,k) \in [T] \times [K] : y_{t,k} \in D^c \right) = 0. \]

Lemma 3.3. Let \( K = \frac{80L^2}{\epsilon^2} \). Given \( t \in [T] \), it holds

\[ E \left[ \|m_{t,k}\| \right] \leq \frac{\epsilon^2}{16}, \]

where \( m_{t,k} = 0 \) for all \( k > k_0 \) if the \( k \)-loop breaks at \( (t,k_0) \). Consequently, for any \( 0 \leq \gamma < 1 \), with probability \( 1 - \gamma \), there are at most \( \log(\gamma^{-1}) \) restarts of the while loop in the \( t \)-th iteration.

We have the following finite-time guarantee for Algorithm 1.

Theorem 3.4. Let \( f \) be \( L \)-Lipschitz continuous. Then, Algorithm 1 with \( K = \frac{80L^2}{\epsilon^2} \) and \( T = \frac{4\Delta}{\epsilon^2} \) finds an \((\epsilon, \delta)\)-GAS point with probability \( 1 - \gamma \) using at most

\[ \frac{320\Delta L^2}{\epsilon^3 \delta} \log \left( \frac{4\Delta}{\gamma \epsilon \delta} \right) \]

oracle calls

with \( P(\epsilon_{\text{error}} = 1) = 0 \), where \( f(x_0) - \inf_x f(x) \leq \Delta \).

3.3. Stochastic Setting

In this section, we consider the stochastic setting. The new algorithm replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(b)) with Assumption 3.1(b).

3.3.1. Algorithm

Technically speaking, the main difference from (Zhang et al., 2020, Algorithm 2) lies in the additional perturbation step.

We need to carefully choose \( \zeta \) to ensure that the iterates are within a \( \delta \)-ball of some reference point without hurting the convergence. Since \( m_t \) is a weighted average of all the stochastic gradients, we need to show that it approximately belongs to the Goldstein \( \delta \)-subdifferential \( \partial_\delta f(x) \) of some reference point \( x \).
The subtlety when \( \|m_t\| = 0 \): Unlike in the deterministic setting where we can terminate the algorithm if \( \|m_{t,k}\| \) is small, in the stochastic case, \( m_t \) is a convex combination of stochastic gradients and thus it does not suffice to terminate the algorithm even if \( \|m_t\| = 0 \). The quantity that we aim to minimize is its expectation \( \mathbb{E}[\|m_t\|] \leq \mathbb{E}[\|m_b\|] \). Due to this subtlety, we cannot let the perturbation size \( \zeta_t \) adapt to \( \|m_t\| \) as in the deterministic case: If \( \zeta_t = \frac{\omega_1 \|m_b\|}{p(\|m_b\| + \omega_2)} \) in Algorithm 2, then when \( \|m_t\| = 0 \), we have \( y_{t+1} = x_{t+1} \), and we cannot ensure that \( f \) is differentiable at \( x_t \) almost surely. We choose a constant \( \zeta_t = \zeta \) in Algorithm 2 instead. In this case, when \( \|m_t\| = 0 \), \( y_{t+1} \) is sampled from a ball centered at \( x_t \).

By adopting a random conic perturbation to (Zhang et al., 2020, Algorithm 2), we have the following Algorithm 2, in which the main differences with (Zhang et al., 2020, Algorithm 2) are marked in blue.

**Algorithm 2 Perturbed Stochastic INGD**

**Input:** \( x_1 \in \mathbb{R}^d \).

**Initialize:** \( m_1 = g_1 = 0 \), \( x_1 \). Set \( \beta = 1 - \frac{c^2}{\epsilon_0^2 \zeta^2} \), \( K = \frac{1}{\ln \frac{\lambda}{\epsilon}} \ln \frac{16G}{\epsilon} \), \( \omega = \left( 1 - \frac{1}{\lambda} \right) \ln \frac{16G}{\epsilon} \), \( p = \frac{64G}{\epsilon} \ln \frac{16G}{\epsilon} \), \( q = \frac{256G^2}{\epsilon^2} \ln \frac{16G}{\epsilon} \), \( T = \frac{2^{16}G^3 \ln \frac{16G}{\epsilon}}{\epsilon^4} \max\{1, \frac{G}{\lambda \epsilon} \} \).

1. Set \( \text{error} = 0 \).
2. for \( t \in [T] \) do
3. \( x_{t+1} = x_t - \eta_t m_t \), where \( \eta_t = \frac{1}{p(\|m_t\| + \gamma)} \).
4. Sample \( u_{t+1} \in \mathbb{R}^{d+1} \) uniformly from \( \mathbb{S}^d \).
5. Let \( v_{t+1} \in \mathbb{R}^d \) be the first \( d \) coordinates of \( u_{t+1} \).
6. If \( \|m_t\| > 0 \), \( b_{t+1} = v_{t+1} \), otherwise, \( b_{t+1} = v_{t+1} \).
7. Sample \( y_{t+1} \) uniformly from \( [x_t, x_{t+1} + \zeta b_{t+1}] \), where \( \zeta = \min\left\{ \frac{p}{\|b_{t+1}\|}, \frac{\epsilon_0}{\epsilon_0 L(1 + \zeta)} \right\} \).
8. Call oracle \( g_{t+1} = \mathcal{O}_{\delta}(y_{t+1}) \).
9. \( m_{t+1} = \beta m_t + (1 - \beta) g_{t+1} \).
10. end for
**Output:** \( x_{\text{out}} := x_{\max\{1, t - K\}} \), where \( i \sim \text{Unif}([T]) \).

### 3.3.2. Finite-Time Analysis

We have the following finite-time guarantee for Algorithm 2, which is similar to (Zhang et al., 2020, Theorem 10) but replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(b)) with Assumption 3.1(b).

**Theorem 3.5.** Under Assumption 3.1(b), with probability at least \( \frac{2}{5} \), the output of Algorithm 2 satisfies \( \text{dist}(0, \partial f(x_{\text{out}})) \leq \epsilon \) after at most

\[
\bar{O}\left( \frac{G^3 \Delta}{\epsilon^4 \delta} \right)
\]

oracle calls

with \( P(\text{error} = 1) = 0 \), where \( f(x_0) - \inf_x f(x) \leq \Delta \).

### 4. Computing NAS by GAS

In this section, we isolate a new function class within which the new algorithmic scheme can compute near-approximately stationary points in finite time. The new class goes far beyond that of \( \rho \)-weakly convex functions. We will first introduce the general results, and then several useful calculus rules. In Section 5, we will discuss applications of the new techniques to modern machine learning models.

#### 4.1. General Results

The main strategy is to compute NAS by GAS. To this end, we need certain continuity of set-valued subdifferential mapping \( \partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \), which should be stronger than upper semicontinuity. A classic notion in set-valued analysis named outer Lipschitz continuity is defined as follows.

**Definition 4.1** (Donchev & Rockafellar (2009, 3D)). A set-valued mapping \( G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) is outer Lipschitz continuous (OLC) at \( y \) relative to a set \( D \) if \( y \in D \subset \text{dom} G \), \( G(y) \) is a closed set, and there is a constant \( \kappa \geq 0 \) along with a neighborhood \( V \) of \( y \) such that

\[
G(y) \subseteq G(y) + \kappa(y - y)B, \quad \forall y \in V \cap D.
\]

OLC is a weaker notion than Lipschitz continuity even for a single-valued mapping \( G : \mathbb{R} \rightrightarrows \mathbb{R} \). See (Lewis & Pang, 2010, Example 2.4(a)). However, for our purposes, OLC is not sufficient since by the classic result of Robinson (1981) the bad function in Proposition 2.7 is OLC.

The following modified OLC notion for set-valued mapping is new and central in our development, which allows us to have a Lipschitz-type control of \( G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) from above within a constant-size neighborhood (see also Figure 2).

**Definition 4.2** (\( (\delta, \eta, \kappa) \)-outer Lipschitz continuous). A set-valued mapping \( G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) is \((\delta, \eta, \kappa)\)-OLC on \( S \) if for any \( x \in S \), there exists a pivot \( y \in B_n(x) \cap S \) such that \( G \) is \( \kappa \)-OLC on \( B_n(x) \cap S \). In other words, for all \( x \in S \), there exists \( a \in B_n(x) \cap S \) such that

\[
G(z) \subseteq G(y) + \kappa(y - z)B, \quad \forall z \in B_n(x) \cap S.
\]

Besides, we call \( P^G : x \rightrightarrows y \) the pivot mapping of \( G \).

**Remark 4.3.** A natural question about Definition 4.2 is why we want to set \( \eta \) and \( \delta \) to different values. In other words, why \((\delta, \eta, \kappa)\)-OLC is not sufficient. Consider the convex function \( g(x, y) := \max\{2x, -2x, y\} \), which is the bad function in the proof of Proposition 2.7. It is easy to see that \( \partial g \) is polyhedral and OLC by (Robinson, 1981). However, for any \( \delta > 0 \), \( \partial g \) is not \((\delta, \delta, \kappa)\)-OLC at \( (0, 2\delta) \), \( \forall \kappa \geq 0 \). Thus, even for an OLC mapping \( \partial g \), we cannot promise \( \exists \delta > 0 \) such that \( \partial g \) is \((\delta, \delta, \kappa)\)-OLC at certain \( x \in \text{dom} \partial g \). Instead, we will show in Theorem 4.5 that if \( \partial g \) is OLC
and $S$ is compact, then $\forall \delta > 0, \exists \eta > 0$ such that $\partial g$ is $(\delta, \eta, \kappa)$-OLC on $S$.

We are now ready for the main theorem of this section:

**Theorem 4.4 (NAS by GAS).** For a Lipschitz continuous $f$, suppose that $\partial f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is $(\delta, \eta, \kappa)$-OLC. If $x$ is $(\epsilon, \eta)$-GAS, then $x$ is $(\epsilon + \kappa(\delta + \eta), \delta)$-NAS.

It is natural to ask what function class admits a $(\delta, \eta, \kappa)$-OLC subdifferential.

**Theorem 4.5.** Let $\delta > 0$ and $\partial f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be $\kappa$-OLC. For any compact set $S$, there exists an $\eta \in (0, \delta]$ such that $\partial f$ is $(\delta, \eta, \kappa)$-OLC on $S$.

**Remark 4.6.** If the set of $(\epsilon, \delta)$-GAS points is bounded and $\partial f$ is OLC, then we can use Theorem 4.5 and Theorem 4.4 to guarantee NAS from GAS. Note that functions with OLC subdifferential have been widely studied in the variational analysis literature. For example, $\partial f$ with a finite union of convex polyhedral graph (Robinson, 1981) is OLC. If $f$ is an $\epsilon$-stable function (Bednarík & Pastor, 2013, Theorem 2), then $\partial f$ is OLC.

Given an OLC mapping $\partial f$ and a constant $\delta > 0$, it is generally hard to estimate the constant $\eta$ as guaranteed by Theorem 4.5. However, its value is needed for the stopping rules of Algorithm 1. In the next subsection, we provide several useful calculus rules to compute the parameter $\eta$ explicitly.

### 4.2. Calculus of $(\delta, \eta, \kappa)$-Outer Lipschitz Continuity

In this section, we establish a series of calculus rules to verify and compute the parameters in Definition 4.2. We first introduce four rules that have taken the subdifferential calculus rules of $f$ into consideration.

**Proposition 4.7 (smooth regularization).** Suppose that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has a $(\delta, \eta, \kappa)$-OLC $\partial f$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable with a $\beta$-Lipschitz gradient $\nabla g$. Then $\partial(f + g)$ is $(\delta, \eta, \beta + \kappa)$-OLC.

**Proposition 4.8 (separable sum).** Suppose, for any $i \in [m]$, that $f_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}$ has a $(\delta_i, \eta_i, \kappa_i)$-OLC $\partial f_i$. Let $f(x) := \sum_{i=1}^m f_i(x_i)$, where $x := \bigoplus_{i=1}^m x_i$. Then, $\partial f$ is $(\delta, \eta, \kappa)$-OLC with

$$\delta = \sum_{i=1}^m \delta_i^2, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i^2.$$

**Proposition 4.9 (linear composition).** Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ has a $(\delta, \eta, \kappa)$-OLC $\partial f$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ is $L_2$-Lipschitz and $\beta$-smooth. Then, $\partial(g \circ f)$ is $(\delta, \eta, \beta L_1 + \kappa L_2)$-OLC.

Then, we introduce a partial sum rule, which is powerful but needs to be used in conjunction with certain subdifferential calculus rules (e.g., assuming Clarke regularity). The following rules are crucial in the 2-layer ReLU neural network example (see Section 5.2.2).

**Proposition 4.11 (sum).** Suppose, for any $i \in [m]$, that $g_i : \mathbb{R}^{d} \Rightarrow \mathbb{R}$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a shared pivot mapping $P : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Let $G(x) := \sum_{i=1}^m g_i(x)$. Then, $G$ is $(\delta, \eta, \kappa)$-OLC with

$$\delta = \min_{i \in [m]} \delta_i, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

**Corollary 4.12 (partially separable sum).** Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d} \Rightarrow \mathbb{R}$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a partially shared pivot mapping $P_i : \mathbb{R}^{d_0} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_0} \times \mathbb{R}^d$, such that $\pi_i \circ P_i(x_0, x_1) = \pi_i \circ P_i(x_0, x_1), \forall x \in [m]$. Let $G(x) := \sum_{i=1}^m G_i(x_0, x_1)$, where $x := \bigoplus_{i=0}^m x_i$. Then, $G$ is $(\delta, \eta, \kappa)$-OLC with

$$\delta = \sum_{i=1}^m \delta_i^2, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

### 4.3. Discussion

We record here a recipe to prove $(\delta, \eta, \kappa)$-OLC from scratch, which when combined with the calculus rules in this section forms a toolbox for determining the parameters $(\delta, \eta, \kappa)$.

1. Construct pivot mapping $P : \mathbb{R}^d \rightarrow \mathbb{R}^d$.
2. Verify $\|x - P(x)\| \leq \delta$ for all $x \in \mathbb{R}^d$.
3. Prove that for all $x \in \mathbb{R}^d$, it holds

$$G(z) \subseteq (P(x) + \kappa \|z - P(x)\|)B, \forall z \in B_{\eta}(x) \cap S.$$

We will provide concrete examples in Section 5.
5. Applications

To demonstrate the wide applicability of the new theoretical framework, we discuss examples in machine learning, namely ρ-margin SVM, 1-layer, and 2-layer ReLU NN, all being Clarke irregular and not weakly convex. We show that all these examples are subdifferential (δ, η, κ)-OLC, where the parameters (δ, η, κ) can be determined via the calculus rules in Section 4.2.

5.1. ρ-Margin loss SVM

We aim to solve

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \phi_\rho(z_i^\top w), \quad (\rho-\text{MSVM})$$

where $\phi_\rho(u) := \left(1, \max \left(0, 1 - \frac{u}{\rho}\right)\right)$.

The goal is to compute ($\epsilon, \delta$)-NAS points for Problem (ρ-MSVM) by computing ($\epsilon', \delta'$)-GAS points. We note that the ρ-Margin loss SVM in Problem (ρ-MSVM) and its ρ = 1 version, also known as ramp loss SVM, have been widely recognized in the operations research literature (Brooks, 2011; Carrizosa et al., 2014; Wang et al., 2021; Tian & So, 2022), statistics (Shen et al., 2003; Wu & Liu, 2007; Liu et al., 2005), and machine learning (Huang et al., 2014; Keshet & McAllester, 2011; Collober et al., 2006; Ertekin et al., 2010; Suzumura et al., 2017; Maibing & Igel, 2015) communities as providing better robustness against data outliers than the vanilla SVM. The general ρ-version can be found in the learning theory textbook (Mohri et al., 2018, Corollary 5.11).

It is elementary to see that $\partial \phi_\rho$ is $(\delta, \delta, 0)$-OLC for any $0 < \delta \leq \frac{\rho}{2}$ with pivot mapping $P^{\phi_\rho} : \mathbb{R} \to \mathbb{R}$ defined by

$$P^{\phi_\rho}(x) := \begin{cases} 0 & \text{for } |x| \leq \frac{\rho}{2}, \\ \rho & \text{for } |x| - \rho < \frac{\rho}{2}, \\ x & \text{for otherwise.} \end{cases}$$

Let $\Phi_\rho(y) := \sum_{i=1}^n \phi_\rho(y_i)$. Then, by Proposition 4.8, $\partial \Phi_\rho$ is $(\sqrt{n}\delta, \delta, 0)$-OLC. Assuming that $Z \in \mathbb{R}^{n \times d}$ is surjective, by Proposition 4.9, $\partial (\Phi_\rho \circ Z)$ is $(\sqrt{n}\delta \|Z\|, \frac{\delta}{\sqrt{2\|Z\|}}, 0)$-OLC. Using Proposition 4.7, $\partial F$ is $(\sqrt{n}\delta \|Z\|, \frac{\delta}{\sqrt{2\|Z\|}}, 1)$-OLC. By Theorem 4.4, if $x$ is $(\epsilon, \frac{\delta}{\sqrt{\|Z\|}})$-GAS, then it is also $(\epsilon + (\sqrt{n}\|Z\| + \frac{1}{\|Z\|}) \delta, \sqrt{n} \|Z\| \|\delta\|)$-NAS. Let the condition number of $Z$ be $\kappa(K) := \|Z\| \|\delta\|$. In other words, to compute an ($\epsilon, \delta$)-NAS point, it is sufficient to have an ($\epsilon', \delta'$)-GAS point, where (in a dimension-free manner)

$$\epsilon' \leq \frac{\epsilon}{2} \text{ and } \delta' \leq \min \left\{ \frac{\delta}{\sqrt{n} \kappa(K)}, \frac{\epsilon}{2\sqrt{n} \kappa(K)} + \frac{\rho}{2\|Z\|} \right\}.$$
OLC trackable, where Vec$(W) \in \mathbb{R}^{nd}$. By $F(W) = f(X_{\text{big}} \text{Vec}(W)) + R(W)$ and Proposition 4.7, $F(W)$ is subdifferentially OLC trackable.

5.2.2. 2-LAYER RELU NEURAL NETWORK

Let $g(a, b) := a \cdot \max\{b, 0\}$. We aim to solve

$$\min_{W \in \mathbb{R}^{n \times d}, a \in \mathbb{R}^n} F(W, a) := \sum_{i=1}^n \left(y_i \cdot \sum_{j=1}^m g(a_j, w_{ij}^T x_i) \right) + R(W, a)$$

with surjective $X \in \mathbb{R}^{n \times d}$ and smooth regularization term $R: \mathbb{R}^{d \times m} \times \mathbb{R}^m \to \mathbb{R}$.

Compared with the 1-layer case, the main difficulty in the analysis is due to the inseparability of $\{a_j\}_{j \in [m]}$, as one cannot apply the subdifferential chain rule and OLC calculus rules directly. To cope with this, we need the partial separable rule in Corollary 4.12 and a partially differentiable sum rule in Proposition 5.2, which might be of independent interest. To begin, we have the following subdifferential characterization of $\partial \varrho : \mathbb{R}^2 \to \mathbb{R}^2$.

**Claim 5.1.** For $g(u_1, u_2) := u_1 \cdot \max\{u_2, 0\}$, it holds

$$\partial \varrho(u_1, u_2) = \begin{cases} (u_2, u_1) & \text{for } u_2 > 0, \\ (0, 0) & \text{for } u_2 < 0, \\ (0, \text{Co}\{0, u_1\}) & \text{for } u_2 = 0. \end{cases}$$

Then, we investigate the continuity of $\partial \varrho$. Given any $\delta > 0$, $x \in \mathbb{R}^2$, and $z \in \mathbb{B}_\delta(x)$, we consider the following cases:

- If $|x_2| > \delta$, let $y = x$.
  - If $y_2 > 0$, then $z_2 > 0$. We have $\partial \varrho(z) = (z_2, z_1) \subseteq (y_2, y_1) + \|y - z\|E = \partial \varrho(y) + \|y - z\|E$.
  - If $y_2 < 0$, then $z_2 < 0$. We have $\partial \varrho(z) = (0, 0) = \partial \varrho(y)$.

- If $0 \leq |x_2| \leq \delta$, let $y = (x_1, 0)$. It is easy to see that $\|y - x\| = |x_2| \leq \delta$.
  - If $z_2 > 0$, we have $\partial \varrho(z) = (z_2, z_1) \subseteq (0, y_1) + \|y - z\|E \subseteq \partial \sigma(y) + \|y - z\|E$.
  - If $z_2 < 0$, we have $\partial \varrho(z) = (0, 0) \subseteq (0, 0) + \|y - z\|E$.
  - If $z_2 = 0$, we have $\partial \varrho(z) = (0, \text{Co}\{0, z_1\}) \subseteq (0, \text{Co}\{0, y_1\}) + \|y - z\|E = \partial \sigma(y) + \|y - z\|E$.

Therefore, for any $\delta > 0$, $\partial \varrho$ is $(\delta, \delta, 1)$-OLC with pivot mapping $P^{\partial \varrho} : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$P^{\partial \varrho}(x, z) := \begin{cases} (x_1, 0) & \text{for } |x| \leq \delta, \\ (x_1, x_2) & \text{otherwise}. \end{cases}$$

It is easy to see that $\pi_1 \circ P^{\partial \varrho}((x_1, x_2))$ is independent of $x_2$. Let $h_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \forall i \in [n]$ be defined by

$$h_i(a, u_i) := \ell_i \left( \sum_{j=1}^m g(a_j, u_{ij}) \right).$$

Then, by the choices of pivots in the proof of Proposition 4.10 and 4.11, $\partial h_i$ is subdifferentially OLC trackable with pivot mapping $P^{\partial h_i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^m$ defined by $P^{\partial h_i}((a, u_i)) := (a, \tilde{u}_i)$, where

$$\tilde{u}_i := \begin{cases} 0 & \text{for } |u_{ij}| \leq \delta, \\ u_{ij} & \text{otherwise}. \end{cases}$$

Therefore, $\{\partial h_i\}_{i \in [m]}$ partially shares the pivot mapping $P^{\partial h_i}$ on the first argument, i.e., $\pi_1 \circ P^{\partial h_i}((x_0, x_i)) = \pi_1 \circ P^{\partial h_i}((y_0, y_i)), \forall i \in [n]$. Let $f(a, U) := \sum_{i=1}^n h_i(a, u_i)$. By Corollary 4.12, $\sum_{i=1}^n \partial h_i$ is subdifferentially OLC trackable. To proceed, we need the following chain rule, whose proof is technical and might be of independent interest.

**Proposition 5.2** (partially differentiable sum rule). It holds

$$\partial f(a, U) = \sum_{i=1}^n \partial h_i(a, u_i).$$

Then, $\partial f$ is subdifferentially OLC trackable. Suppose that the data $X \in \mathbb{R}^{n \times d}$ is surjective. Let $x_i$ be the $i$-th row of $X$. We define $\theta \in \mathbb{R}^{m+nd}$ and $X_{\text{huge}} \in \mathbb{R}^{(m+nd) \times (m+nd)}$ as

$$\theta := \left[ \begin{array}{c} a \\ \text{Vec}(W) \end{array} \right], \quad X_{\text{huge}} := \left[ \begin{array}{c} I_m \\ X_{\text{big}} \end{array} \right].$$

As $X$ is surjective, $X_{\text{huge}}$ is surjective. Using Proposition 4.9, we have $f(X_{\text{huge}}, \theta)$ is subdifferentially OLC trackable. By $F(W, a) = f(X_{\text{huge}}, \theta) + R(W, a)$ and Proposition 4.7, $F(W, a)$ is subdifferentially OLC trackable.

6. Closing Remarks

In this paper, we report a practical algorithmic scheme to compute GAS points for general Lipschitz functions with finite-time complexity. We also isolate a new function class for which our scheme computes NAS points in finite time. Besides, we establish a series of theoretical tools to compute parameters in our new function class. To demonstrate the wide applicability of our new theoretical framework, we discuss modern machine learning models and show that they satisfy our new conditions. We hope that our results can be beneficial to the understanding of finite-time complexity of sharper approximate stationarity for Lipschitz continuous “non”-problems. An intriguing further direction is to apply the new analytical framework to other nonconvex nonsmooth problems. Extending the calculus rules in Section 4.2 or refining the modified OLC notion in Definition 4.2 would also be interesting.
References


A. Proofs of Section 2

Figure 3. The function used in the proof of Proposition 2.7.

**Proposition 2.7** (convex polyhedron). For any $\delta > 0$, there exists a convex function $f : \mathbb{R}^2 \to \mathbb{R}$, which is 2-Lipschitz with polyhedral $\partial f$, such that $(0, 2\delta)$ is $(0, \delta)$-GAS but $\min_{y \in B_{\delta}((0, 2\delta))} \text{dist}(0, \partial f(y)) \geq \frac{2}{5}\sqrt{5}$.

**Proof.** Fixing some $\delta > 0$, consider the function (see also Figure 3), whose convexity is obvious,

$$f(x, y) := \max\{2x, -2x, y\}.$$ 

Note that, by $(\pm \delta, 2\delta) \in B_{\delta}((0, 2\delta))$, it holds

$$(0, 0) = \frac{1}{2}(-2, 0) + \frac{1}{2}(2, 0) \in \frac{1}{2}\partial f((-\delta, 2\delta)) + \frac{1}{2}\partial f((\delta, 2\delta)) \subseteq \partial f((0, 2\delta)).$$

Besides, as $(0, 0) \notin B_{\delta}((0, 2\delta))$, it is elementary to see

$$\text{dist}(0, \bigcup_{y \in B_{\delta}((0, 2\delta))} \partial f(y)) \geq \min_{0 \leq \lambda \leq 1} \|(2\lambda, 0) + (0, 1 - \lambda)\| = \frac{2}{5}\sqrt{5},$$

as required. \hfill $\square$

B. Proofs of Section 3.2

**Lemma 3.2.** Let $D := \{x : f \text{ is differentiable at } x\}$. Given locally Lipschitz continuous $f$, we have

$$\mathbb{P}\left(\exists (t, k) \in [T] \times [K] : y_{t, k} \in D^c\right) = 0.$$

**Proof.** Fix $(t, k) \in [T] \times [K]$. Let

$$S_1 := \{ (\lambda, \xi) : \lambda \in [0, 1], \xi \in \mathbb{R}^{d-1}, \|\xi\| \leq 1 \},$$

$$S_2 := \left\{ y \in \mathbb{R}^d : y = x_t + \lambda \left( x_{t, k} - x_t + \frac{\delta\|m_{t, k}\|}{8L} \cdot b_{t, k+1} \right), \lambda \in [0, 1], \|b_{t, k+1}\| \leq 1, b_{t, k+1}^T(x_{t, k} - x_t) = 0 \right\}.$$

Let $X^\perp \in \mathbb{R}^{d \times d-1}$ be an orthonormal basis of span$(x_{t, k} - x_t)^\perp$. We define the following isomorphism:

$$T : S_1 \to S_2$$

$$(\lambda, \xi) \to y_{t, k+1} := x_t + \lambda \left( x_{t, k} - x_t + \frac{\delta\|m_{t, k}\|}{8L} \cdot X^\perp \xi \right).$$

Then, by Rademacher theorem (Rockafellar & Wets, 2009, Theorem 9.60) and $T^{-1}$ is Lipschitz, we have

$$m(\{y \in D^c \cap S_2\} = m(\{ (\lambda, \xi) \in T^{-1}(D^c \cap S_2) \}) = 0.$$
Let $S_3 := \{ b \in \mathbb{R}^d : \| b \| \leq 1, b^T (x_{t,k} - x_t) = 0 \}$. By (Barthe et al., 2005, Corollary 4), we have $b_{t,k+1} \sim \text{Unif}(S_3) \overset{d}{=} X^\perp \xi$, where $\xi \sim \text{Unif}(B^{d-1})$. With $\lambda \sim \text{Unif}([0, 1])$ and countable union of zero measure set is negligible, we have

$$P(y_{t,k} \in D^c \cap S_2, \forall (t,k) \in [T] \times [K]) = 0,$$

which completes the proof.

**Lemma 3.3.** Let $K = \frac{8g\lambda^2}{\epsilon^2}$. Given $t \in [T]$, it holds

$$\mathbb{E}\left[ \left\| m_{t,K}\right\|^2 \right] \leq \frac{\epsilon^2}{16},$$

where $m_{t,k} = 0$ for all $k > k_0$ if the $k$-loop breaks at $(t,k_0)$. Consequently, for any $0 \leq \gamma < 1$, with probability $1 - \gamma$, there are at most $\log(\gamma^{-1})$ restarts of the while loop in the $t$-th iteration.

**Proof.** Let $F_{t,k} = \sigma(y_{t,1}, \ldots, y_{t,k})$ and $\tilde{F}_{t,k} = \sigma(y_{t,1}, \ldots, y_{t,k}, b_{t,k+1})$. We denote $D_{t,k}$ as the event that $k$-loop does not break at $x_{t,k}$, i.e., $\| m_{t,k} \| > \epsilon$ and $f(x_{t,k}) - f(x_t) > -\frac{\delta}{4}\| m_{t,k} \|$. It is clear that $D_{t,k} \in F_{t,k} \subset \tilde{F}_{t,k}$. Let

$$\gamma(\lambda) = (1 - \lambda)x_t + \lambda (x_{t,k} + \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1})$$

for $\lambda \in [0, 1]$. Note that $\gamma'(\lambda) = x_{t,k} - x_t + \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}$. Let

$$x'_{t,k} = x_{t,k} + \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}.$$ 

Since $y_{t,k+1}$ is uniformly sampled from the line segment $[x_t, x'_{t,k}]$ and $f$ is differentiable at $y_{t,k+1}$ almost surely by Lemma 3.2, we know that

$$\mathbb{E}\left[ \langle g_{t,k+1}, x'_{t,k} - x_t \rangle \| \tilde{F}_{t,k} \right] = \int_0^1 f'(\gamma(t); x'_{t,k} - x_t) dt = f(x'_{t,k}) - f(x_t).$$

By $x'_{t,k} - x_t = -(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}$, we have

$$\mathbb{E}\left[ \langle g_{t,k+1}, m_{t,k} \rangle \| \tilde{F}_{t,k} \right]
= -\frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}} \cdot \mathbb{E}\left[ \langle g_{t,k+1}, x'_{t,k} - x_t \rangle \| \tilde{F}_{t,k} \right] + \frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}} \cdot \mathbb{E}\left[ \langle g_{t,k+1}, \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1} \rangle \| \tilde{F}_{t,k} \right]
\leq -\frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}} \cdot \left( f(x_{t,k}) - f(x_t) - |f(x'_{t,k}) - f(x_{t,k})| \right) + \frac{\| m_{t,k} \|^2}{8(1 - \frac{\| m_{t,k} \|}{8L})}
\leq -\frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \frac{\delta\| m_{t,k} \|}{8L} \cdot b_{t,k+1}} \cdot \left( f(x_{t,k}) - f(x_t) \right) + \frac{\| m_{t,k} \|^2}{4(1 - \frac{\| m_{t,k} \|}{8L})},$$

which directly implies

$$\mathbb{E}\left[ \langle g_{t,k+1}, m_{t,k} \rangle \| F_{t,k} \right] \leq -\frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \delta \left( f(x_{t,k}) - f(x_t) \right) + \frac{\| m_{t,k} \|^2}{4(1 - \frac{\| m_{t,k} \|}{8L})}}.$$ 

By construction, $m_{t,k+1} = \beta m_{t,k} + (1 - \beta) g_{t,k+1}$ under $D_{t,k} \cap \cdots \cap D_{t,1}$, and $m_{t,k+1} = 0$ otherwise. Let $D_{t,k} = D_{t,k} \cap \cdots \cap D_{t,1}$. Therefore,

$$\mathbb{E}\left[ \| m_{t,k+1} \|^2 \| F_{t,k} \right]
\leq \left( \beta^2 \| m_{t,k} \|^2 + (1 - \beta)^2 L^2 + 2\beta(1 - \beta) \cdot \mathbb{E}\left[ \langle g_{t,k+1}, m_{t,k} \rangle \| F_{t,k} \right] \mathbf{1}_{D_{t,k}} \right)
\leq \left( \beta^2 \| m_{t,k} \|^2 + (1 - \beta)^2 L^2 + 2\beta(1 - \beta) \cdot \left( -\frac{\| m_{t,k} \|}{(1 - \frac{\| m_{t,k} \|}{8L}) \cdot \delta \left( f(x_{t,k}) - f(x_t) \right) + \frac{\| m_{t,k} \|^2}{4(1 - \frac{\| m_{t,k} \|}{8L})}} \right) \mathbf{1}_{D_{t,k}} \right)
\leq \left( \beta^2 \| m_{t,k} \|^2 + (1 - \beta)^2 L^2 + 2\beta(1 - \beta) \cdot \frac{\| m_{t,k} \|^2}{2(1 - \frac{\| m_{t,k} \|}{8L})} \right) \mathbf{1}_{D_{t,k}} =: h(\beta) \mathbf{1}_{D_{t,k}}.$$
By rearranging, it holds
\[
    h(\beta) = \beta^2 \left( L^2 + \|m_{t,k}\|^2 - \frac{\|m_{t,k}\|^2}{1 - \frac{\|m_{t,k}\|}{8L}} \right) + \beta \left( \frac{\|m_{t,k}\|^2}{1 - \frac{\|m_{t,k}\|}{8L}} - 2L^2 \right) + L^2.
\]

Note that, by $0 < \|m_{t,k}\| \leq L$ in $D_{t,k}$, and conditioning on $D_{t,k}$, it holds
\[
    T_1 = \frac{L^2}{8(1 - \frac{\|m_{t,k}\|}{8L})} \left( 8 - \frac{\|m_{t,k}\|}{L} - \frac{\|m_{t,k}\|^3}{L^3} \right) > 0, \quad \text{and} \quad T_2 = \frac{L^2}{4(1 - \frac{\|m_{t,k}\|}{8L})} \left( -8 + \frac{\|m_{t,k}\|}{L} + 4\|m_{t,k}\|^2 \right) < 0.
\]

Thus, $h(\beta)$ achieves the minimum at $\beta_{t,k} = \frac{8L^3 - 8L^2\|m_{t,k}\| - 4L\|m_{t,k}\|^2}{64L^4 - 16L^3\|m_{t,k}\| + L^2\|m_{t,k}\|^2 - 8L\|m_{t,k}\|^3 + \|m_{t,k}\|^4}$.

Let $0 < t := \frac{\|m_{t,k}\|}{L} \leq 1$. For the validity of inequality $(\sharp)$, we observe
\[
    \frac{1}{5} < \inf_{0 < t \leq 1} \frac{16 - 8t + t^2}{64 - 16t + t^2 - 8t^3 + t^4} \leq c_1.
\]

To see it, note that, for $0 < t \leq 1$, it holds
\[
    5 \cdot (16 - 8t + t^2) - (64 - 16t + t^2 - 8t^3 + t^4) = (t + 2) \left( 10 \left( t - \frac{4}{5} \right)^2 + \frac{8}{5} - t^3 \right) > 0.
\]

Therefore,
\[
    \mathbb{E} \left[ \|m_{t,k+1}\|^2 \right] = \mathbb{E} \mathbb{E} \left[ \|m_{t,k+1}\|^2 | \mathcal{F}_{t,k} \right] \leq \mathbb{E} \left[ \left( 1 - \frac{\|m_{t,k}\|^2}{5L^2} \right) \|m_{t,k}\|^2 \right] \leq \left( 1 - \frac{\mathbb{E}[\|m_{t,k}\|^2]}{5L^2} \right) \mathbb{E}[\|m_{t,k}\|^2].
\]

Then, by a similar argument in the proof of (Zhang et al., 2020, Lemma 13) we have $\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{5L^2}{K + t^1}$. When $K \geq \frac{80L^2}{\epsilon}$, we have $\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{\epsilon^2}{16}$. □

**Theorem 3.4.** Let $f$ be $L$-Lipschitz continuous. Then, Algorithm 1 with $K = \frac{80L^2}{\epsilon^2}$ and $T = \frac{4\Delta}{\epsilon^2}$ finds an $(\epsilon, \delta)$-GAS point with probability $1 - \gamma$ using at most
\[
    \frac{320\Delta L^2}{\epsilon^3 \delta} \log \left( \frac{4\Delta}{\gamma \epsilon^2} \right)
\]
oracle calls

with $P(\text{error} = 1) = 0$, where $f(x_0) - \inf_x f(x) \leq \Delta$.

**Proof.** Using Lemma 3.2 and Lemma 3.3, the remaining parts directly follow from the proof of (Zhang et al., 2020, Theorem 8). □

**C. Proofs of Section 3.3**

**Theorem 3.5.** Under Assumption 3.1(b), with probability at least $\frac{3}{5}$, the output of Algorithm 2 satisfies $\text{dist}(0, \partial_h f(x_{out})) \leq \epsilon$ after at most
\[
    \tilde{O} \left( \frac{G^3 \Delta}{\epsilon^3 \delta} \right)
\]
oracle calls

with $P(\text{error} = 1) = 0$, where $f(x_0) - \inf_x f(x) \leq \Delta$. □
Proof. Let $\alpha_i := \beta^{t-i}(1 - \beta)$ and denote $x'_{i+1} := x_{i+1} + \zeta b_{i+1}$ and $G_t := \sigma(g_1, \ldots, g_t), \forall t$. Clearly, the random variables $m_t, x_t, x_{t+1}, \eta_t$ are $G_t$-measurable. Note that

$$m_t = \beta^K m_{t-K} + \sum_{i=t-K+1}^{t} \alpha_i g_i.$$  

Conceptually, if we choose $K$ to be sufficiently large, the term $\beta^K m_{t-K}$ is negligible. Then, if all the points $y_{t-K+1}, \ldots, y_t$ are inside $x_{t-K} + \delta B$, we have that $m_t$ approximately belongs to $\partial_\delta f(x_{t-K})$ in expectation.

Note that for all $i = t - K + 1, \ldots, t$,

$$\|y_i - x_{t-K}\| \leq \|y_i - x_{i-1}\| + \|x_{i-1} - x_{t-K}\|$$

(a) holds since $y_i$ is sampled from the line segment $[x_{i-1}, x'_i]$, (b) uses $\|b_k\| \leq 1$ and (c) follows from $\zeta \leq \frac{\omega}{p}$ and $\eta_t \|m_i\| \leq \frac{1}{p}, \forall t$. We verify that the choices of $K, \omega$ and $p$ satisfy $K + \omega \leq \delta$:

$$\frac{K + \omega}{p} \leq \frac{1 - \beta}{\beta} \ln \frac{16G}{\delta e^\epsilon} = \delta.$$

Then, conditioned on $G_{t-K}$, since for all $i = t - K + 1, \ldots, t$,

$$\mathbb{E}[g_i | G_{t-K}] = \mathbb{E}[\nabla f(y_i) | G_{t-K}] \in \partial_\delta f(x_{t-K}),$$

we have (note that $\sum_{i=t-K+1}^{t} \alpha_i = 1 - \beta^K$)

$$\frac{1}{1 - \beta^K} \sum_{i=t-K+1}^{t} \alpha_i \mathbb{E}[g_i | G_{t-K}] \in \partial_\delta f(x_{t-K})$$

$\Rightarrow \frac{1}{1 - \beta^K} (\mathbb{E}[m_t | G_{t-K}] - \beta^K m_{t-K}) \in \partial_\delta f(x_{t-K})$

$\Rightarrow \text{dist}(0, \partial_\delta f(x_{t-K})) \leq \frac{1}{1 - \beta^K} \left( \| \mathbb{E}[m_t | G_{t-K}] \| + \beta^K \|m_{t-K}\| \right)$

$\leq \frac{1}{1 - \beta^K} \left( \mathbb{E}[\|m_t\| | G_{t-K}] + \beta^K \|m_{t-K}\| \right).$

Take expectation.

$$\mathbb{E}[\text{dist}(0, \partial_\delta f(x_{t-K}))] \leq \frac{1}{1 - \beta^K} \mathbb{E}[\|m_t\|] + \frac{\beta^K G}{1 - \beta^K},$$

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\text{dist}(0, \partial_\delta f(x_{t-K}))] \leq \frac{1}{(1 - \beta^K)T} \sum_{t=1}^{T} \mathbb{E}[\|m_t\|] + \frac{\beta^K G}{1 - \beta^K}.$$

Practical Optimization of Lipschitz Functions with Finite-Time Complexity
We verify that the choices of $\beta$ and $K$ satisfy $\beta^kG \leq \frac{\epsilon}{16}$: $(\beta^k \leq \frac{\epsilon}{16G})$ $\iff$ $(K \geq \frac{1}{\ln \frac{16G}{\epsilon}})$. WLOG, we assume that $\epsilon \leq G$, and thus $\beta^k \leq \frac{1}{16}$. The above inequality can be further bounded as

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [\text{dist}(0, \partial x_t f(x_{t-K}))] \leq \frac{16}{15T} \sum_{t=1}^{T} \mathbb{E} [\|m_t\|] + \epsilon \frac{1}{15}. \quad (C.1)$$

The remaining proof is to show that Algorithm 2 ensures that $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [\|m_t\|] = O(\epsilon)$.

For ease of analysis, we denote $\tilde{Y}_{t+1} := \sigma(g_1, \ldots, g_t, b_{t+1}, y_{t+1})$ and $\tilde{Y}_{t+1} := \sigma(g_1, \ldots, g_t, b_{t+1})$. Clearly, we have $\mathcal{G}_t \subseteq \tilde{Y}_{t+1} \subseteq \tilde{Y}_{t+1} \subseteq \mathcal{G}_{t+1}$. Let $\varphi(\lambda):=(1-\lambda)x_t + \lambda x'_{t+1}$ for $\lambda \in [0,1]$. Since $y_{t+1}$ is uniformly sampled from the line segment $[x_t, x'_{t+1}]$ and that $f$ is differentiable at $y_{t+1}$ almost surely, it holds that

$$\mathbb{E} [\langle g_{t+1}, x'_{t+1} - x_t \rangle | \mathcal{G}_t] = \mathbb{E} [\mathbb{E} [\langle g_{t+1}, x'_{t+1} - x_t \rangle | \tilde{Y}_{t+1}] | \mathcal{G}_t] = \mathbb{E} [\mathbb{E} [\nabla f(y_{t+1}, x'_{t+1} - x_t) | \tilde{Y}_{t+1}] | \mathcal{G}_t] = \mathbb{E} [\int_0^1 f'(\varphi(\lambda); x'_{t+1} - x_t) d\lambda | \mathcal{G}_t] = \mathbb{E} [f(x'_{t+1}) - f(x_t) | \mathcal{G}_t]. \quad (C.2)$$

By $x'_{t+1} - x_t = -\eta_t m_t + \zeta b_{t+1}$, we have

$$\mathbb{E} [\langle g_{t+1}, x'_{t+1} - x_t \rangle | \mathcal{G}_t] = -\eta_t \mathbb{E} [\langle g_{t+1}, m_t \rangle | \mathcal{G}_t] + \zeta \mathbb{E} [\langle g_{t+1}, b_{t+1} \rangle | \mathcal{G}_t] \leq -\eta_t \mathbb{E} [\langle g_{t+1}, m_t \rangle | \mathcal{G}_t] + \zeta G,$n

where we used $\|b_{t+1}\| \leq 1$. Thus, combining with (C.2), we obtain

$$\mathbb{E} [\langle g_{t+1}, m_t \rangle | \mathcal{G}_t] \leq \frac{1}{\eta_t} \mathbb{E} [f(x_t) - f(x_{t+1}) + f(x_{t+1}) - f(x'_{t+1}) | \mathcal{G}_t] + \frac{\zeta}{\eta_t} G \leq \frac{1}{\eta_t} (f(x_t) - f(x_{t+1})) + \frac{\zeta}{\eta_t} (L + G). \quad (C.3)$$

Based on the construction $m_{t+1} = \beta m_t + (1 - \beta) g_{t+1}$, we can conclude that

$$\mathbb{E} \left[ \eta_t \left( \|m_{t+1}\|^2 - \beta^2 \|m_t\|^2 \right) \right] = 2 \beta (1 - \beta) \mathbb{E} [\eta_t (\langle g_{t+1}, m_t \rangle)] + (1 - \beta)^2 \mathbb{E} [\|g_{t+1}\|^2].$$

From (C.3), it holds that

$$\mathbb{E} \left[ \eta_t \left( \|m_{t+1}\|^2 - \beta^2 \|m_t\|^2 \right) \right] \leq 2 \beta (1 - \beta) \mathbb{E} [f(x_t) - f(x_{t+1})] + 2 \beta (1 - \beta) (L + G) \zeta + (1 - \beta)^2 \mathbb{E} \left[ \eta_t \|g_{t+1}\|^2 \right] + \frac{2 \beta (1 - \beta) \Delta}{T} + 2 \beta (1 - \beta) (L + G) \zeta + \frac{(1 - \beta)^2 G^2}{q},$$

where we used $\eta_t \leq \frac{1}{q}$.

Since $\eta_t = \frac{1}{p \|m_t\| + q}$, using the same telescoping proof in (Zhang et al., 2020), as long as $\frac{2G}{q} \leq \frac{\beta}{2}$, the following holds

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \eta_t \left( \|m_{t+1}\|^2 - \beta^2 \|m_t\|^2 \right) \right] \geq \frac{\beta (1 - \beta)}{2T} \sum_{t=1}^{T+1} \mathbb{E} \left[ \frac{\|m_t\|^2}{p \|m_t\| + q} - \frac{\beta G^2}{qT} \right].$$
Thus, 
\[
\frac{\beta(1 - \beta)}{2T} \sum_{t=1}^{T+1} \mathbb{E} \left[ \frac{\|m_t\|^2}{p \|m_t\| + q} \right] \leq \frac{2\beta(1 - \beta)\Delta}{T} + 2\beta(1 - \beta)(L + G)\zeta + \frac{(1 - \beta)^2G^2}{q} + \frac{\beta G^2}{qT},
\]
\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \frac{q \|m_t\|^2}{p \|m_t\| + q} \right] \leq \frac{4q\Delta}{T} + 4q(L + G)\zeta + \frac{2(1 - \beta)G^2}{\beta} + \frac{2G^2}{T(1 - \beta)}.
\]

Comparing the above inequality with (14)\(^5\) in (Zhang et al., 2020), we notice that the only difference is the additional perturbation term \(4q(L + G)\zeta\). Since we choose the identical \(\beta, p, q\) and \(T\) as in (Zhang et al., 2020), using the arguments (15) and (16) in (Zhang et al., 2020) and denoting \(m_{\text{avg}} := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|m_t\|\), we obtain
\[
\frac{4Gm_{\text{avg}}^2}{m_{\text{avg}} + 4G} \leq \frac{\epsilon^2}{17} + 4q(L + G)\zeta
\]
\[
\leq \frac{\epsilon^2}{15},
\]
where \((*)\) uses \(\zeta \leq \frac{\epsilon^2}{50q(L + G)}\). The above is a quadratic equation in \(m_{\text{avg}}\):
\[
4Gm_{\text{avg}}^2 - \frac{\epsilon^2}{15}m_{\text{avg}} - \frac{4G\epsilon^2}{15} \leq 0.
\]
Solving for the positive root of this quadratic equation and using \(\epsilon \leq G\), we obtain
\[
m_{\text{avg}} \leq \frac{\epsilon^2}{15} + \sqrt{\frac{\epsilon^4}{225} + \frac{64G^2\epsilon^2}{15}} \leq \frac{4\epsilon}{15} \leq \frac{5\epsilon}{16}.
\]

Finally, using (C.1), we conclude that
\[
\mathbb{E} \left[ \text{dist}(0, \partial_{\delta}f(x_{\text{out}})) \right] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \text{dist}(0, \partial_{\delta}f(x_{t-\delta})) \right] \leq \frac{2\epsilon}{5}.
\]

Thus, with probability at least \(\frac{3}{4}\), we have \(\text{dist}(0, \partial_{\delta}f(x_{\text{out}})) \leq \epsilon\). \(\square\)

**D. Proofs of Section 4.1**

**Theorem 4.4 (NAS by GAS).** For a Lipschitz continuous \(f\), suppose that \(\partial f : \mathbb{R}^d \rightarrow \mathbb{R}^d\) is \((\delta, \eta, \kappa)\)-OLC. If \(x\) is \((\epsilon, \eta)\)-GAS, then \(x\) is \((\epsilon + \kappa(\delta + \eta), \delta)\)-NAS.

**Proof.** As \(x\) is Goldstein \((\epsilon, \eta)\)-stationary, we have \(\text{dist}(0, \partial f(x + \eta B)) \leq \epsilon\), which implies that there exists
\[
\|g\| \leq \epsilon, \quad \text{such that} \quad g \in \partial f(x + \eta B) = \text{Co} \left\{ \bigcup_{y \in B_{\eta}(x)} \partial f(y) \right\}.
\]

By Carathéodory’s theorem (Rockafellar & Wets, 2009, Theorem 2.29), we can write \(g = \sum_{j=1}^{d+1} \alpha_j g_j\), where \(\alpha_j \geq 0\), \(\sum_{j=1}^{d+1} \alpha_j = 1\), \(g_j \in \partial f(y_j), y_j \in B_{\eta}(x), \forall j \in [d + 1]\).

Let \(y \in B_{\delta}(x)\) be a pivot such that \(\partial f\) is \(\kappa\)-outer Lipschitz continuous on \(B_{\eta}(x)\). As \(f\) is Lipschitz and by (Clarke, 1990, Proposition 2.1.2), \(\partial f(y)\) is nonempty, convex, and compact. Let \(u_j := \arg \min_{z \in \partial f(y)} \left\| z - g_j \right\|, u := \sum_{j=1}^{d+1} \alpha_j u_j \in \partial f(y)\).

Then, we compute
\[
\|u\| = \left\| \sum_{j=1}^{d+1} \alpha_j u_j \right\| \leq \left\| g \right\| + \kappa \sum_{j=1}^{d+1} \alpha_j \|y_j - y_j\| \leq \left\| g \right\| + \kappa \sum_{j=1}^{d+1} \alpha_j \left(\|y_j - x\| + \|x - y_j\|\right) \leq \epsilon + \kappa(\delta + \eta),
\]
\(^5\)There is a typo in the telescoping proof of Theorem 14 in (Zhang et al., 2020): The term \(\frac{\alpha^2 G^2}{q}\) above Equation (14) should be \(\frac{\beta G^2}{q}\). This typo does not affect the final convergence result.
which completes the proof. □

**Theorem 4.5.** Let $\delta > 0$ and $\partial f : \mathbb{R}^d \rightarrow \mathbb{R}$ be $\kappa$-OLC. For any compact set $S$, there exists an $\eta \in (0, \delta]$ such that $\partial f$ is $(\delta, \eta, \kappa)$-OLC on $S$.

**Proof.** Let $\bigcup_{x \in S} B^\circ_{\frac{\delta}{2}}(x) (x)$ be an open cover of $S$, where $\delta(x) = \min \{\delta, \delta(x)\}$ and $\delta(x)$ is the inradius of neighborhood $V(x)$, on which $\partial f$ is $\kappa$-Lipschitz at $x$, satisfying $B^\circ_{\delta(x)}(x) \subseteq V(x)$. As $S$ is compact, we find a finite subcover $\bigcup_{i \in [m]} B^\circ_{\frac{\delta}{2}}(x_i)$, where $\delta_i = \min \{\delta, \delta(x_i)\}$ and $x_i \in S$. Let $\eta := \min_{i \in [m]} \frac{\delta}{2}$. Then, by Lebesgue’s number theorem (Munkres, 1974, Chapter 3, Lemma 7.2) on open cover $\bigcup_{x \in S} B^\circ_{\delta(x)}(x)$ of $S$, for any $x \in S$, there exists $i \in [m]$ such that $B^\circ_{\eta}(x) \subseteq B^\circ_{\eta}(x_i)$. Thus, $B_{\eta}(x) \subseteq B^\circ_{\delta}(x_i) \subseteq V(x_i)$. For any $z \in B_{\eta}(x) \cap S$, by $\kappa$-outer Lipschitz continuity $\partial f$ on $V(x_i)$, we have

$$\partial f(z) \leq \partial f(x_i) + \kappa\|z - z\|_B, \quad \forall z \in B_{\eta}(x) \cap S,$$

where $x_i \in B^\circ_{\delta}(x) \subseteq B^\circ_{\delta}(x)$. This completes the proof. □

### E. Proofs of Section 4.2

**Proposition 4.7** (smooth regularization). Suppose that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has a $(\delta, \eta, \kappa)$-OLC $\partial f$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable with a $\beta$-Lipschitz gradient $\nabla g$. Then $\partial(f + g)$ is $(\delta, \eta, \beta + \kappa)$-OLC.

**Proof.** Let $F := f + g$. By (Rockafellar & Wets, 2009, Exercise 8.8(c)), $\partial F = \partial f + \nabla g$. Let $y \in B_{\delta}(x)$ be a pivot of $\partial f$. Then, for $\forall z \in B_{\eta}(x)$, we compute

$$\nabla F(z) = \partial f(z) + \nabla g(z) \leq \partial f(z) \leq \partial f(y) + \beta\|z - y\|_B \leq \partial F(y) + (\beta + \kappa)\|z - y\|_B,$$

which completes the proof. □

**Proposition 4.8** (separable sum). Suppose, for any $i \in [m]$, that $f_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}$ has a $(\delta_i, \eta_i, \kappa_i)$-OLC $\partial f_i$. Let $f(x) := \sum_{i=1}^m f_i(x_i)$, where $x := \bigoplus_{i=1}^m x_i$. Then, $\partial f$ is $(\delta, \eta, \kappa)$-OLC with

$$\delta = \sqrt{\sum_{i=1}^m \delta_i^2}, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sqrt{\sum_{i=1}^m \kappa_i^2}.$$

**Proof.** By (Rockafellar & Wets, 2009, Proposition 10.5) and $f$ is Lipschitz, $\partial f = \bigoplus_{i=1}^m \partial f_i$. Let $y_i \in B^\circ_{\delta_i}(x_i)$ be a pivot of $\partial f_i$. Also $y := \bigoplus_{i=1}^m y_i$. Similarly, for any $z \in B_{\eta}(x)$, it holds $z_i \in B^\circ_{\eta_i}(x_i) \subseteq B^\circ_{\eta_i}(x_i), \forall i \in [m]$. We compute

$$\partial f(z) = \bigoplus_{i=1}^m \partial f_i(z_i) \subseteq \bigoplus_{i=1}^m \left( \partial f_i(y_i) + \kappa_i\|y_i - z_i\|_{B^{d_i}} \right) \subseteq \partial f(y) + \kappa\|z - y\|_B^d,$$

where $\|y - x\|^2 = \sum_{i=1}^m \|y_i - x_i\|^2 \leq \sum_{i=1}^m \delta_i^2 = \delta^2$. This completes the proof. □

**Proposition 4.9** (linear composition). Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ has a $(\delta, \eta, \kappa)$-OLC $\partial f$ and $A \in \mathbb{R}^{n \times d}$ is surjective. Then, $\partial(f \circ A)$ is $(\delta\|A^\dagger\|, \frac{n}{\|A\|}, \kappa\|A\|^2)$-outer Lipschitz continuous.

**Proof.** Let $F(x) := f(Ax)$. As $A$ is surjective, by (Rockafellar & Wets, 2009, Exercise 10.7), $\partial F(x) = A^\dagger \partial f(Ax)$. Let $q \in B^\circ_\eta(Ax)$ be a pivot of $\partial f$. Let $y := A^\dagger q + (I - A^\dagger A)x$. Then $Ay = q$ and $\|y - x\| \leq \|A\|\|q - Ax\| \leq \delta\|A\|^2$. Meanwhile, for any $z \in B_{\eta}(x)$, it holds $\|Ax - Ax\| \leq \delta\|A\|^2 \|x - y\| \leq \eta$. We compute

$$\partial F(z) = A^\dagger \partial f(Az) \subseteq A^\dagger \partial f(Ay) + \kappa\|Ay - Az\|A^\dagger B^n \subseteq \partial F(y) + \kappa\|A\|^2\|y - z\|_B^d,$$

which completes the proof. □

**Proposition 4.10** (rescaling). Suppose that the $L_1$-Lipschitz $f : \mathbb{R}^n \rightarrow \mathbb{R}$ has a $(\delta, \eta, \kappa)$-OLC $\partial f$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ is $L_2$-Lipschitz and $\beta$-smooth. Then, $\partial(g \circ f)$ is $(\delta, \eta, \beta L_1 + \kappa L_2)$-OLC.
Proof. Let $F = g \circ f$. By (Clarke, 1990, Theorem 2.3.9(ii)), $\partial F(x) = \nabla g(f(x)) \cdot \partial f(x)$. Let $y \in B_\delta(x)$ be a pivot of $\partial f$. Then, for $\forall z \in B_\eta(x)$, we compute

$$
\partial F(z) = \nabla g(f(z)) \cdot \partial f(z) \subseteq \nabla g(f(z)) \cdot \partial f(y) + \nabla g(f(z)) \cdot \kappa \|z - y\|B
$$

$$
\subseteq \nabla g(f(y)) \cdot \partial f(y) + (\beta L_1 + \kappa L_2)\|z - y\|B
$$

$$
= \partial F(y) + (\beta L_1 + \kappa L_2)\|z - y\|B,
$$

which completes the proof. \qed

Proposition 4.11 (sum). Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^d \supseteq \mathbb{R}^d$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a shared pivot mapping $P : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Let $G(x) := \sum_{i=1}^{m} G_i(x)$. Then, $G$ is $(\delta, \eta, \kappa)$-OLC with

$$
\delta = \min_{i \in [m]} \delta_i, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^{m} \kappa_i.
$$

Proof. Let $y = P(x) \in B_\delta(x)$ be a pivot of $G_i(x)$, which by pivot sharing assumption should hold for all $i \in [m]$. Thus $\|y - x\| \leq \min_{i \in [m]} \delta_i = \delta$. For all $z \in B_\eta(x) \subseteq B_\eta(x)$, we compute

$$
G(z) = \sum_{i=1}^{m} G_i(z) \subseteq \sum_{i=1}^{m} \left( G_i(y) + \kappa_i \|z - y\|B \right) \subseteq G(y) + \left( \sum_{i=1}^{m} \kappa_i \right) \|z - y\|B,
$$

as expected. \qed

Corollary 4.12 (partially separable sum). Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \supseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a partially shared pivot mapping $P_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$, such that $\pi_1 \circ P_i(x_0, x_i) = \pi_1 \circ P_i(x_0, x_1)$, $\forall i \in [m]$. Let $G(x) := \sum_{i=1}^{m} G_i(x_0, x_i)$, where $x := \bigoplus_{i=0}^{m} x_i$. Then, $G$ is $(\delta, \eta, \kappa)$-OLC with

$$
\delta = \sqrt{\sum_{i=1}^{m} \delta_i^2}, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^{m} \kappa_i.
$$

Proof. Let $(y_0, y_i) \in B^{d_0+d_i}(x_0, x_i)$ be a pivot of $G_i$. Also $y := \bigoplus_{i=0}^{m} y_i$. Similarly, for any $z \in B_\eta(x)$, it holds $(z_0, z_i) \in B^{d_0+d_i}(x_0, x_i) \subseteq B^{d_0+d_i}(x_0, x_i), \forall i \in [m]$. We compute

$$
G(z) = \sum_{i=1}^{m} G_i((z_0, z_i)) \subseteq \sum_{i=1}^{m} \left( G_i((y_0, y_i)) + \kappa_i \|(z_0, z_i) - (y_0, y_i)\|B^{d_0+d_i} \right) \subseteq G(y) + \kappa \|z - y\|B^d,
$$

where $\|y - x\|^2 = \sum_{i=0}^{m} \|y_i - x_i\|^2 \leq \sum_{i=1}^{m} (\|y_0 - x_0\|^2 + \|y_i - x_i\|^2) \leq \sum_{i=1}^{m} \delta_i^2 = \delta^2$, and $d = \sum_{i=0}^{m} d_i$. This completes the proof. \qed

F. Proofs of Section 5

Claim 5.1. For $g(u_1, u_2) := u_1 \cdot \max\{u_2, 0\}$, it holds

$$
\partial g(u_1, u_2) = \begin{cases} 
(u_2, u_1) & \text{for } u_2 > 0, \\
(0, 0) & \text{for } u_2 < 0, \\
(0, \text{Co}\{0, u_1\}) & \text{for } u_2 = 0.
\end{cases}
$$

Proof. Define

$$
C_1 := \{(u_1, u_2) : u_2 \geq 0\}, \\
C_2 := \{(u_1, u_2) : u_2 \leq 0\}.
$$
It is clear that $C_1 \cup C_2 = \mathbb{R}^2$, and we have
\[
\varrho(u_1, u_2) = \begin{cases} 
    u_1 \cdot u_2 & \text{for } (u_1, u_2) \in C_1, \\
    0 & \text{for } (u_1, u_2) \in C_2.
\end{cases}
\]

Note that $C_1 \cap C_2$ form a set $S$ of measure 0, and if $(u_1, u_2) \notin S$, then $\varrho$ is differentiable. The claim follows from taking convex hull with (Rockafellar & Wets, 2009, Theorem 9.61). \hfill \Box

In the following proof, we will use the following notion named partial Clarke subdifferential. See also (Clarke, 1990, Page 48), (Rockafellar & Wets, 2009, Corollary 10.11).

**Definition F.1.** Let a local Lipschitz function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $g_y : x \to f(x, y)$. Then the partial Clarke subdifferential with respect to the first argument is defined as $\partial_1 f(x, y) := \partial g_y(x)$. $\partial_2 f(x, y)$ is defined similarly.

**Claim F.2.** $\partial \varrho(u_1, u_2) = \partial_1 \varrho(u_1, u_2) \times \partial_2 \varrho(u_1, u_2)$ and $|\pi_1 \circ \partial \varrho(u_1, u_2)| = 1$.

**Proof.** Note that $\partial_1 \varrho(u_1, u_2) = \max \{u_2, 0\}$ and $\partial_2 \varrho(u_1, u_2) = u_1 \cdot \partial (\max \{\cdot, 0\})(u_2)$. The proof completes by using Claim 5.1 and literally checking definitions. \hfill \Box

**Proposition 5.2** (partially differentiable sum rule). It holds
\[
\partial f(a, U) = \sum_{i=1}^n \partial h_i(a, u_i).
\]

**Proof.** To begin, we observe the following general fact. For any set $A \subseteq \mathbb{R}^n \times \mathbb{R}^m$, if $|\pi_1 A| = 1$, then $A = \pi_1 A \times \pi_2 A$. To see it, for one direction, if $(a_1, a_2) \in A$, then $a_1 \in \pi_1 A, a_2 \in \pi_2 A$. Thus, $A \subseteq \pi_1 A \times \pi_2 A$. For the other direction, let $a_1 \in \pi_1 A, a_2 \in \pi_2 A$. As $\{a_1\} = \pi_1 A$, then by the definition of $\pi_2 A$, it holds $(a_1, a_2) \in A$. Thus $\pi_1 A \times \pi_2 A \subseteq A$.

To avoid uninformative sophisticated notation, we will use “$\equiv$” for equivalence up to coordinate permutation. Formally, if $A \subseteq B$, then there exists a permutation matrix $P$ such that $B = \{Px : x \in A\}$. We compute
\[
\partial f(a, U) \equiv \sum_{i=1}^n \partial h_i(a, u_i) \quad \text{(Clarke (1990, Proposition 2.3.3))} \\
= \sum_{i=1}^n \nabla \ell_i \left( \sum_{j=1}^m \varrho(a_j, u_{ij}) \right) \cdot \sum_{j=1}^m \partial \varrho(a_j, u_{ij}) \quad \text{(Clarke (1990, Theorem 2.3.9(ii)))} \\
= \sum_{i=1}^n \nabla \ell_i \left( \sum_{j=1}^m \varrho(a_j, u_{ij}) \right) \cdot \sum_{j=1}^m \left( \partial_1 \varrho(a_j, u_{ij}) \times \partial_2 \varrho(a_j, u_{ij}) \right) \quad \text{(Claim F.2)} \\
= P \left( \sum_{i=1}^n \nabla \ell_i \left( \sum_{j=1}^m \varrho(a_j, u_{ij}) \right) \cdot \sum_{j=1}^m \partial_1 \varrho(a_j, u_{ij}) \right) \times \sum_{i=1}^n \nabla \ell_i \left( \sum_{j=1}^m \varrho(a_j, u_{ij}) \right) \cdot \sum_{j=1}^m \partial_2 \varrho(a_j, u_{ij}) \right). \\
\]

Note that $|S_1| = 1$ as $|\partial_1 \varrho(a_j, u_{ij})| = 1$, $\forall (i, j) \in [n] \times [m]$ by Claim F.2. Thus $|\pi_1 \circ \partial f(a, U)| = 1$ and $\partial f(a, U) \equiv \pi_1 \circ \partial f(a, U) \times \pi_2 \circ \partial f(a, U)$. With (Clarke, 1990, Proposition 2.3.16), we compute
\[
\partial f(a, U) \equiv \left( \pi_1 \circ \partial f(a, U) \right) \times \left( \pi_2 \circ \partial f(a, U) \right) \\
\supseteq \partial_1 f(a, U) \times \partial_2 f(a, U) \\
\equiv S_1 \times S_2.
\]

To see (z), note that $f(\cdot, U)$ is differentiable. Thus it is straightforward to check $S_1 = \partial_1 f(a, U)$. For $S_2$, note that $f(a, \cdot)$ is fully separable (as $a$ is fixed). Then, with (Rockafellar & Wets, 2009, Proposition 10.5) and (Clarke, 1990, Theorem 2.3.9(ii)), the verification of $S_2 = \partial_2 f(a, U)$ is routine.

This completes the proof. \hfill \Box